Computer Vision and Image Understanding 124 (2014) 110–122

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

# Recommendations for recognizing video events by concept vocabularies

Amirhossein Habibian *, Cees G.M. Snoek

*Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*

ARTICLE INFO

*Article history:*
Received 18 August 2013
Accepted 4 February 2014

*Keywords:*
Event recognition
Concept representation
Concept vocabulary

ABSTRACT

Representing videos using vocabularies composed of concept detectors appears promising for generic event recognition. While many have recently shown the benefits of concept vocabularies for recognition, studying the characteristics of a universal concept vocabulary suited for representing events is ignored. In this paper, we study how to create an effective vocabulary for arbitrary-event recognition in web video. We consider five research questions related to the number, the type, the specificity, the quality and the normalization of the detectors in concept vocabularies. A rigorous experimental protocol using a pool of 1346 concept detectors trained on publicly available annotations, two large arbitrary web video datasets and a common event recognition pipeline allow us to analyze the performance of various concept vocabulary definitions. From the analysis we arrive at the recommendation that for effective event recognition the concept vocabulary should *(i)* contain more than 200 concepts, *(ii)* be diverse by covering *object*, *action*, *scene*, *people*, *animal* and *attribute* concepts, *(iii)* include both general and specific concepts, *(iv)* increase the number of concepts rather than improve the quality of the individual detectors, and *(v)* contain detectors that are appropriately normalized. We consider the recommendations for recognizing video events by concept vocabularies the most important contribution of the paper, as they provide guidelines for future work.



## 1. Introduction

We consider the problem of recognizing events in arbitrary web video, such as the ones depicted in Fig. 1. Among the many challenges involved, resulting from the uncontrolled recording condition of web videos and the large variations in the visual appearance of events, probably one of the most fundamental questions in event recognition is what defines an event in video? The Oxford English dictionary defines an event as "anything that happens". With such a broad definition it is not surprising that the topic has been addressed in the computer vision and multimedia retrieval community by many researchers from diverse angles [4,55,44,5,25,54,36].

In this paper, we study representations that contribute to defining events for automatic recognition. We are inspired by findings from cognition, where research has repeatedly shown that humans remember events by their actors, actions, objects, and locations [46]. Studying event representation based on such high-level concepts is now within reach because of the continued progress in supervised concept detection [48] and the availability of labeled training collections like the ones developed in benchmarks such as TRECVID [47], ImageNet [7] and several other venues [34,11]. Different from concepts, which represent a single person, object, scene or action in videos, events are commonly defined as a more complex interaction of several persons, objects, and actions happening in a specific scene [31]. In this paper, we name the set of available concept detectors as the *vocabulary* and we study how to construct a vocabulary suited for effective recognition of events in video.

The state-of-the-art in event recognition represents a video in terms of low-level audiovisual features [16,38,50,35,15,19,37]. In general, these methods first extract from the video various types of static and/or dynamic features, *e.g.,* color SIFT variations [53], MFCC [15], and Dense Trajectories [38]. Second, the descriptors are quantized and aggregated [38]. The robustness and efficiency of various low-level features for recognizing events are evaluated in [50,15,33]. Despite their good recognition performance, especially when combined together [35,15,38,33], low-level features are incapable of providing an understanding of the semantic structure present in an event. Hence, it is not easy to derive how these event definitions arrive at their recognition. Therefore, essentially different representations are needed for events. We focus on high-level representations for event recognition.

Inspired by previous work in object recognition [51,22], scene recognition [22,40] and activity recognition [41], many have explored high-level representations for recognition of events

* Corresponding author.
*E-mail addresses:* a.habibian@uva.nl (A. Habibian), cgmsnoek@uva.nl (C.G.M. Snoek).

http://dx.doi.org/10.1016/j.cviu.2014.02.003
1077-3142/

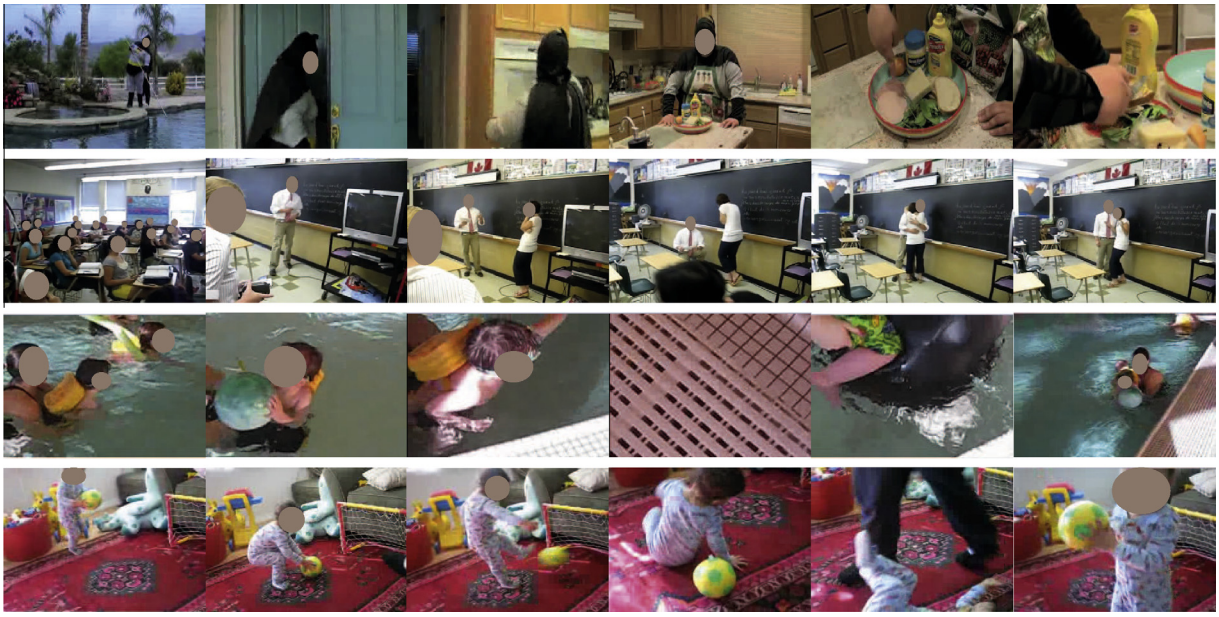


**Fig. 1.** Examples of web videos depicting events. From top to bottom: "*making a sandwich*", "*marriage proposal*", "*swimming*" and "*soccer*". Each event can be defined by its key concepts including actor, place, action and the involved objects.

[28,31,2,58,14,30,8,24]. All these works follow a general pipeline consisting of three consecutive steps to arrive at a high-level video representation. First, frame extraction, where the video is decoded and a subset of frames is extracted. Second, concept detection, where each extracted frame is represented by a vector of predictions from vocabulary concept detectors. Finally, video pooling, where the frame representations are averaged and aggregated into the video level representation. The obtained high-level representation is not only semantically interpretable, but is also reported to outperform the state-of-the-art low-level audiovisual features in recognizing events [31,33]. Rather than training vocabulary concept detectors and event detectors separately, recent work aims for jointly learning the vocabulary concept and event detectors [26,57,1,27]. In these works, the vocabulary concept detectors are trained to optimize the event detection, without explicitly optimizing the individual concept detector accuracy. As a consequence, the vocabulary concepts do not necessarily have a semantic interpretation needed to explain the video content. In this paper, we follow [31,14,58,30] and train concept and event detectors separately.

Identifying a universal vocabulary of concepts suited for representing events is an important question that has been ignored in the literature. To the best of our knowledge, all the previous work on high-level representations for event recognition relies on an arbitrary set of concepts as the vocabulary. By contrast, we focus in this paper on characterizing the vocabulary which is most effective for representing events. We investigate the concept vocabulary from two perspectives: first by characterizing the composition, where we investigate *what* concepts should be included in the vocabulary. Second by characterizing the concept detectors, where we study *how* to create vocabulary concept detectors that are most suited for representing events. Before detailing our research questions, we discuss related work that we consider most relevant to these two perspectives.

## 2. Related work

### 2.1. Characterizing concept vocabulary composition

Our study is inspired by the pioneering work of Hauptmann et al. [9] who focus on construction of concept vocabularies for broadcast news video retrieval. They examined how big the concept vocabulary should be and what concepts should be part of the vocabulary for effective shot retrieval. For this purpose, they used a pool of concepts to create and evaluate vocabularies under different circumstances. In their work, the presence and absence of 320 human-annotated concepts was used as the main source for the investigations. To make the experiments more realistic they insert noise into the human annotations to simulate the behavior of automatic concept detectors. They concluded that 5000 detectors with modest quality would be sufficient for general-purpose broadcast news video retrieval. Regarding the important question what concepts to include in the vocabulary, Hauptmann et al. [9] conclude that frequent concepts contribute more to overall news video retrieval performance than rare concepts, so they are preferred to be included in the vocabulary. However, it is not clear whether their conclusion generalizes to *event* recognition on the challenging domain of unconstrained web video.

In this paper, we start from the analysis by Hauptmann et al. [9] and adopt their research questions for event recognition. Our work is different with respect to the following five aspects. First, we focus exclusively on events, whereas [9] considers news use cases like *Find shots of US Maps depicting the electoral vote distribution (blue vs. red state)* and *Find shots of Refugee Camps with women and children visible.* Second, our domain of study is unconstrained web video, rather than the highly structured broadcast television domain. Third, we place special emphasis on the importance of various concept types in representing events (*e.g.,* objects, scenes, actions, etc.), rather than considering all concepts equally important. Fourth, we evaluate retrieval accuracy on video-level rather than shot-level. Finally, in our analysis we do not rely on human concept annotations directly, but instead we use real detector predictions with varying levels of accuracy per concept. Using the real detectors to represent videos leads to surprising new findings, as we will show in the experiments.

### 2.2. Characterizing vocabulary concept detectors

Automatic detection of concepts in videos is a well studied topic in computer vision and multimedia for which many algorithms have been proposed [49,52,10,17]. These include descriptors, *e.g.,*

SIFT variations [53] and STIP [20], descriptor quantization strategies, *e.g.,* Bag-of-Words, VLAD [13] and Fisher vector coding [42], the use of spatial pyramids [21] and various types of kernels to train classifiers, *e.g.,* RBF, $\chi^2$ and Histogram Intersection [29,59]. Choosing among these options provides us with a wide range of concept detectors with varying accuracies and computational costs. In this paper, we investigate how to create vocabulary concept detectors that are most suited for representing events by considering detectors of varying accuracy.

The state-of-the-art in video concept detection employs an SVM classifier to train detectors. SVM predictions are real-valued numbers that could be positive or negative. To perform the subsequent processing steps on the prediction scores, they should be normalized. The general SVM normalization approach in the literature [31,58,14] is to fit a sigmoid function on top of the prediction scores to estimate the posterior probabilities of concept presence [39,23]. The sigmoid function parameters are estimated from a held out partition of the concept detectors training data. In case the training and test data distributions differ, a common scenario when using pre-trained concept detectors for event recognition in arbitrary video, the detector reliability suffers [56]. Hence, the normalization should be executed with care. In this paper, we examine the influence of normalizing the predictions of vocabulary concept detectors on video event recognition accuracy. For this purpose, we consider several existing score normalizations [12].

We consider supervised normalization, which relies on labeled training data to fit the normalization function, *e.g.,* Sigmoid normalization, and unsupervised normalization that does not require any labeled training data. To circumvent supervision, some unsupervised normalizations make assumptions about the distribution of scores. Z-score normalization, for example, assumes the scores have a Gaussian distribution, so the scores are normalized by shifting and scaling by their mean and standard deviation [12]. Others do not make any assumption about the distribution of scores. For example the recent W-score normalization, which models the tails of score distributions by the Extreme Value Theory [43] from statistics and then uses the models to estimate the concept presence probabilities. We assess the influence of normalizing detectors in a concept vocabulary for event recognition.

### 2.3. Research questions

Our study on the effectiveness of concept vocabularies for video event recognition, is directed by five research questions. The first three questions investigate the ideal concept vocabulary composition, while the last two questions consider the creation of the vocabulary concept detectors. A preliminary version of this study has been published in [8]. Here we put more emphasis on characterizing the vocabulary concept detectors. Our five research questions are:

**RQ1** *How many concepts to include in the vocabulary?*
**RQ2** *What concept types to include in the vocabulary?*
**RQ3** *Which concepts to include in the vocabulary?*
**RQ4** *How accurate should the concept detectors be?*
**RQ5** *How to normalize the concept detectors?*

As humans remember events by the high level concepts they contain, *viz.,* actors, actions, objects, and locations [46], studying the characteristics of the concepts that humans use to describe events could be inspirational for automated event recognition. Therefore, before describing our experimental protocol to address the research questions, we first study the vocabulary that humans use to describe events in videos.

## 3. Human event description

To analyze the vocabulary that humans use to describe events, we utilize a set of textual descriptions written by humans to describe web videos containing events. We process textual descriptions for 13,265 videos, as provided by the TRECVID 2012 Multimedia Event Detection task corpus [47]. For each web video in this corpus a textual description is provided that summarizes the event happening in the video by highlighting its dominant concepts. Fig. 2 illustrates some videos and their corresponding textual descriptions.

After removing stop words and stemming, we end up with 5433 distinct terms from the 13,265 descriptions making up a human vocabulary for describing events. Naturally, the frequency of these terms varies, as also observed by [9]. Most of the terms seldom occur in event descriptions. Whereas, only a few terms have high term-frequencies. To be precise, 50% of the terms occur once in the descriptions and only 2% occurs more than five times. Terms like `man`, `girl`, `perform` and `street` appear most frequent, while `bluefish`, `conductor`, `Mississippi` and `Bulgarian` are instances of less frequent terms. Looking into the vocabulary, we observe that the terms used in human descriptions can be mapped to five distinct concept types, as typically used in the computer vision and multimedia literature: *objects*, *actions*, *scenes*, *visual attributes* and *non-visual concepts*. We manually assign each vocabulary term into one of these five types. After this exercise we observe that 44% of the terms refer to *objects*. Moreover, we note that a considerable number of objects are dedicated to various types of *animals* and *people*; *i.e.*, `lion`, and `teen`. About 21% of the terms depict *actions*, like `walking`. Approximately 10% of the concept types are about *scenes*, such as `kitchen`. *Visual attributes* cover about 13% of the terms; *i.e.*, `white`, `flat`, and `dirty`. The remaining 12% of the terms belong to concepts, which are *not visually depictable*; *i.e.*, `poem`, `problem`, and `language`. We summarize the statistics of our human event descriptions in Fig. 3.

We observe that when describing video events, humans use terms with varying generalizations. Some terms are very specialized and refer to specific objects; like, `salmon`, `cheesecake` and `sand castle`. While other terms are more general and refer to a broader set of concepts; like `human`, `vegetation` and `outdoor`. We analyze the generalization of the vocabulary terms using their depth in the WordNet hierarchy [32]. In this hierarchy, the terms are structured based on their hypernym/hyponym relations, so the more specialized terms are placed at the deeper levels. Our study shows that the 5433 vocabulary terms have an average depth of $9.07 \pm 5.29$. The high variance in term depths indicates that the human vocabulary to describe events is composed of both specific and general terms.

To summarize, analyzing the available event descriptions, we observe that the vocabulary that humans use to describe events is composed of a few thousand words, derived from five distinct concept types: *objects*, *actions*, *scenes*, *visual attributes* and *non-visual concepts*. Moreover, we observe that the vocabulary contains both specific and general concepts. Strengthened by these observations about the human vocabulary for describing events, we design five experiments to answer our research questions on the ideal vocabulary for recognizing events in arbitrary web video.

## 4. Experimental setup

To answer the research questions raised in Section 2.3, we create a rigorous empirical setting. First, we introduce the video datasets used to evaluate the event recognition experiments. Then we explain the pool of concept detectors, which we employ to create


**Fig. 2.** Examples of videos and human-added textual descriptions, from which we study how humans describe events.

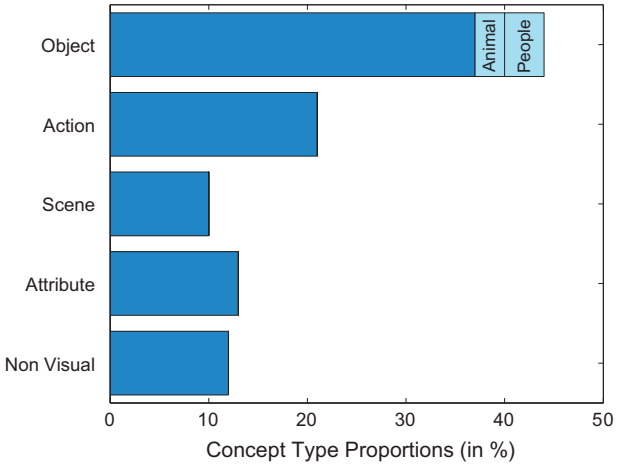



**Fig. 3.** We divide the human vocabulary for describing events into concept types containing *objects*, *actions*, *scenes*, *attributes* and *non-visual* concepts. Our analysis reveals that objects and actions constitute 65% of the human vocabulary when describing events.

vocabularies. Finally, the pipeline used for event recognition using concept vocabularies is presented.

### 4.1. Video datasets

For the event recognition experiments, we rely on two publicly available video collections: the TRECVID Multimedia Event Detection [47] and the Columbia Consumer Video [18] datasets. **TRECVID MED** [47] consists of 13,274 event videos sampled from the TRECVID 2012 Multimedia Event Detection task corpus, as used in [8]. This dataset consists of over 400 h of user-generated video with a large variation in quality, length and content. Moreover, it comes with ground-truth annotations at video level for 25 real-world events, including life events, instructional events, sport events, etc. Following [8], the dataset is divided into a training set (66%) and a test set (34%).

**Columbia CV** [18] contains 9317 user-generated videos from YouTube. This dataset consists of over 210 h of videos in total, where each video has an average length of 80 s. Moreover, the dataset contains ground-truth annotations at video level for 20 semantic categories, where 15 of them are events. The other five categories are objects and scenes, which are excluded from the dataset in our experiments: *"bird"*, *"cat"*, *"dog"*, *"beach"* and *"playground"*. We use the training and test set divisions as defined in [18].

We summarize the training and test set statistics for both video datasets per event in Table 1.

### 4.2. Concept vocabulary

To create the vocabularies, we need a comprehensive pool of concept detectors. We build this pool of detectors using the human annotated training data from two publicly available resources: the TRECVID 2012 Semantic Indexing task [47,3] and the ImageNet Large-Scale Visual Recognition Challenge 2011 [6]. The former has annotations for 346 semantic concepts on 400,000 keyframes from web videos. The latter has annotations for 1000 semantic concepts on 1,300,000 photos. The categories are quite diverse and include concepts from various types; *i.e.*, *object*, *scene* and *action*. Note that the training data is different from the TRECVID

**Table 1**
Number of positive videos in the datasets used in our experiments, split per event. The number of negative videos for each event are around 8800 and 4500 for the TRECVID MED and the Columbia CV dataset, respectively.

| TRECVID MED | | | Columbia CV | | |
|---|---|---|---|---|---|
| Event | Train | Test | Event | Train | Test |
| Attempting board trick | 98 | 49 | Basketball | 182 | 181 |
| Feeding animal | 75 | 48 | Baseball | 150 | 151 |
| Landing fish | 71 | 36 | Soccer | 161 | 162 |
| Wedding ceremony | 69 | 35 | Ice skating | 192 | 193 |
| Working wood working project | 79 | 40 | Skiing | 197 | 196 |
| Birthday party | 121 | 61 | Swimming | 199 | 202 |
| Changing vehicle tire | 75 | 37 | Biking | 136 | 137 |
| Flash mob gathering | 115 | 58 | Graduation | 143 | 145 |
| Getting vehicle unstuck | 85 | 43 | Birthday | 158 | 160 |
| Grooming animal | 91 | 46 | Wedding reception | 129 | 130 |
| Making sandwich | 83 | 42 | Wedding ceremony | 111 | 110 |
| Parade | 105 | 50 | Wedding dance | 174 | 176 |
| Parkour | 75 | 38 | Music performance | 403 | 403 |
| Repairing appliance | 85 | 43 | Non-music performance | 345 | 346 |
| Working sewing project | 86 | 43 | Parade | 191 | 194 |
| Attempting bike trick | 43 | 22 | | | |
| Cleaning appliance | 43 | 22 | | | |
| Dog show | 43 | 22 | | | |
| Giving directions location | 43 | 22 | | | |
| Marriage proposal | 43 | 22 | | | |
| Renovating home | 43 | 22 | | | |
| Rock climbing | 43 | 22 | | | |
| Town hall meeting | 43 | 22 | | | |
| Winning race without vehicle | 43 | 22 | | | |
| Working metal crafts project | 43 | 22 | | | |

MED videos and their textual descriptions, which are used for studying the human vocabulary, as discussed in Section 3.

Leveraging the annotated data available in these datasets, we train 1346 concept detectors in total. We follow the state-of-the-art for our implementation of the concept detectors. We use densely sampled SIFT, OpponentSIFT and C-SIFT descriptors [53] with Fisher vector coding [42]. The codebook used has a size of 256 words. As a spatial pyramid we use the full image and three horizontal bars [21]. The feature vectors representing the training images form the input for a fast linear Support Vector Machine [45].

As summarized in Fig. 3, the concepts that humans use to describe events are derived from *object*, *action*, *scene*, *attributes* and *non-visual* concept types. It is hard to imagine that *non-visual* concepts can be detected by their visual features, so we exclude them from our study. With respect to the importance of the actors in depicting events [46], as well as their high frequency in human descriptions, we consider *people* and *animal* as extra concept types in our experiments. Inspired by this composition, we divide our concept pool by manually assigning each concept to one of the six types. Consequently, we end up with the following concept types: *object* containing 706 concepts, *action* containing 36 concepts, *scene* containing 135 concepts, *people* containing 83 concepts, *animal* containing 338 concepts and *attribute* containing 48 concepts. Fig. 4 provides an overview of the concept types and shows example instances.

### 4.3. Event recognition

In the event recognition experiments, we follow the common pipeline as used in the literature [31,58,14,30]. Unless noted otherwise we use the following implementation. We decode the videos by uniformly extracting one frame every two seconds. Then all the concept detectors are applied on the extracted frames. After concatenating the detector outputs, each frame is represented by a concept vector. Finally, the frame representations are pooled into a video level representation by averaging and normalizing as proposed in [43]. On top of this concept vocabulary representation per video, we use again a linear SVM classifier to train the event recognizers.

## 5. Experiments

We perform five experiments to address our research questions. Each concept vocabulary used in the experiments is evaluated based on its performance in recognizing events using the datasets, pipeline and evaluation protocol described in Section 4. Moreover, the vocabularies are all derived from the concept pool introduced in Section 4.2.

- **Experiment 1: How many concepts to include in the vocabulary?** To study this question, we create several vocabularies with varying sizes and evaluate their performance for recognizing events. Each vocabulary is made of a random subset of the concept detectors from the concept pool. To compensate for possible random effects, all experiments are repeated 50 times and the results are averaged.
- **Experiment 2: What concept types to include in the vocabulary?** We look into this question by comparing two types of vocabularies: *(i) single type* vocabularies, where all concepts are derived from one type and *(ii) joint type* vocabularies, where concepts are derived from all available concept types. We perform this experiment for six kinds of single type vocabularies: *object*, *action*, *scene*, *people*, *animal* and *attribute* types respectively.
  To make the single type and joint type vocabularies more comparable, we force the vocabularies to be of equal size. We do so by randomly selecting the same number of concepts from the concept pool. All the experiments are repeated 500 times to balance possible random effects.
- **Experiment 3: Which concepts to include in the vocabulary?** In this experiment, we investigate whether the concept vocabulary for event recognition should be made of general concepts, specific concepts, or their mixture. We manually label and select two sets of general and specific concepts from the concept pool. The former contains 149 general concepts, *i.e.*,

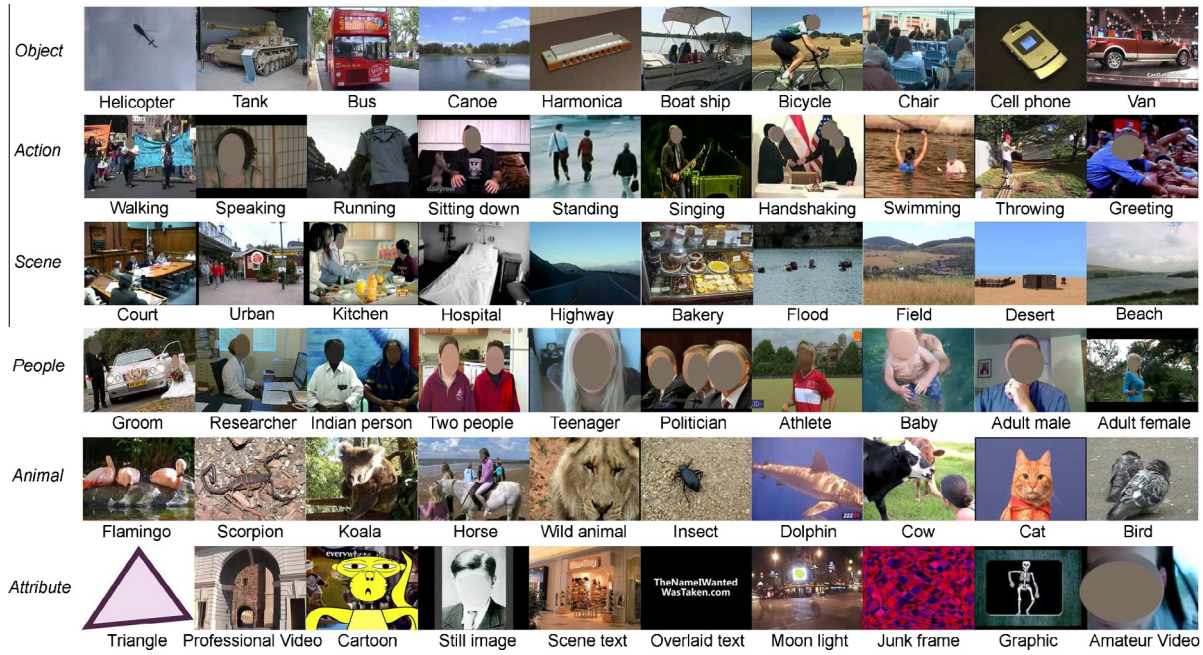



**Fig. 4.** Random training examples of the 1346 concept detectors included in the overall vocabulary used in our experiments, grouped by their concept type.

`vegetation`, `human` and `man made thing`, and the latter contains 619 specific concepts, *i.e.*, `religious figure`, `emergency vehicle` and `pickup truck`. The rest of the concepts, which are not clearly general or specific, are not involved in this experiment. Using these sets we compare three types of vocabularies: *(i)* a *general* vocabulary in which all the concepts are general, *(ii)* a *specific* vocabulary in which all the concepts are specific and *(iii)* a *mixture* vocabulary in which the concepts are randomly selected from both general and specific concept sets. We repeated this experiment for different vocabulary sizes and found that the results remained stable. The reported results are obtained for a vocabulary size of 70, averaged over 500 repetitions.

- **Experiment 4: How accurate should the concept detectors be?** We look into this question by decreasing the detector accuracies and measuring how the event recognition performance responds. To decrease the detector accuracies we follow two different approaches: the first approach trains less sophisticated concept detectors, and the second approach imposes noise into the concept prediction scores.
  In the first approach, we train four versions of our vocabulary concept detectors at different levels of sophistication: *(i) 100%-3SIFT-SP* is the most sophisticated version, where the detectors are implemented as described in Section 4.2. In this version, detectors are trained on all available training data. *(ii) 30%-3SIFT-SP* is similar to *100%-3SIFT-SP*, but the detectors are trained on a random subset of 30% of the available concept training examples. *(iii) 30%-SIFT-SP* is similar to *30%-3SIFT-SP*, but does not include any color SIFT and only relies on standard intensity SIFT. *(iv) 30%-SIFT* is the same as *30%-SIFT-SP*, but without using any spatial pyramid. The four versions of the detectors are trained for the 346 semantic concepts from the TRECVID Semantic Indexing dataset.
  In the second approach, we make the concept detectors inaccurate by gradually imposing increasing amounts of noise into their predictions. The output of each concept detector, as an SVM classifier, is a real-valued number which is supposed to be larger than +1 and smaller than −1 for positive and negative samples. However in practice, the SVM only assigns these values to the samples which are confidently classified, while other samples are assigned to the unconfident area in between −1 and 1. Looking into the concept detector predictions, we observe that most of them are agglomerated in the unconfident area. The less accurate a concept detector is, the more samples are assigned to the unconfident area. To simulate the detector accuracy changes, we randomly select predictions and shift them towards the center of the unconfident area, which has the least decision confidence. We gradually increase the amount of noise and repeat the experiments 50 times to compensate for possible random factors.
- **Experiment 5: How to normalize the concept detectors?** In this experiment, we investigate the effect of normalizing concept vocabularies on video event recognition accuracy. We compare the representation obtained from un-normalized predictions with the representations obtained by applying several normalizations as introduced in Section 2.2: supervised, unsupervised, assumption-based and assumption-free normalization. We apply sigmoid normalization as a supervised method and compare it with Z-score and W-score, as instances of unsupervised normalizations. Moreover, to study the effect of making assumptions on the distribution of concept detector predictions we compare Z-score, which assumes a Gaussian distribution, with W-score normalization, which is an assumption-free method.

Each experiment results in a ranking of the videos from both the test sets based on the probability that the video contains the event of interest. As the evaluation criterion for these ranked lists, we employ average precision (AP) which is in wide use for evaluating visual retrieval results [47]. We also report the average performance over all events as the mean average precision (MAP).

## 6. Results

### 6.1. Experiment 1: How many?

As shown in Fig. 5, adding more concept detectors to the vocabulary improves the event recognition performance. The

improvement gain is particularly prevalent for small vocabularies. When increasing the vocabulary from 50 to 300, on TRECVID MED for example, the MAP increases from 0.125 to 0.221. The improvement is less prevalent when more than 1000 detectors are part of the vocabulary. When increasing the vocabulary from 1000 to 1346 the absolute MAP improvement is only 0.012 on average. We observe similar behavior on Columbia CV. We speculate that the improvement comes from the finer gain partitioning of the event feature space, which in our case is caused by the concept annotations, but is also achievable along other means [57].

However, by looking into individual event recognition results we observe that not all events behave similar when increasing the vocabulary size. For some events, *i.e.*, *"flash mob gathering"*, a relatively high average precision of 0.34 is obtained by including only 50 concepts. We observe that there are some concepts within the vocabulary which are very discriminative for this event *i.e.*, `group of people`, `dancing` and `people marching`. In contrast, for some other events *i.e.*, *"giving directions to a location"*, the event recognition performance is not improved by increasing the vocabulary size. Apparently, there is no concept in the vocabulary, which can effectively discriminate this event from others. It demonstrates that besides the vocabulary size, the relevance of vocabulary concepts should be considered.

The error bars plotted in Fig. 5 indicate the variance in MAPs for various vocabularies. The variance demonstrates that with the same number of concept detectors, some vocabularies perform better than others. In the next two experiments, we study the characteristics of these optimal vocabularies.

Small vocabularies have poor performances in recognizing events. In addition, their effectiveness could be rapidly increased by adding a few more concept detectors. So, in general we recommend to include at least 200 concept detectors in the vocabulary.

### 6.2. Experiment 2: What concept types?

Tables 2 and 3 compare single type and joint type vocabularies for recognizing events. Comparing the MAPs for both datasets, we conclude that joint type vocabularies outperform single type vocabularies for all six concept types on average. It demonstrates that when creating the vocabulary, it is better to sample the concept detectors from diverse types. Hence, we need to detect the objects, people, actions and scenes occurring in the video *jointly* to recognize the event properly. In other words, all of the concept types contribute to the recognition of events.

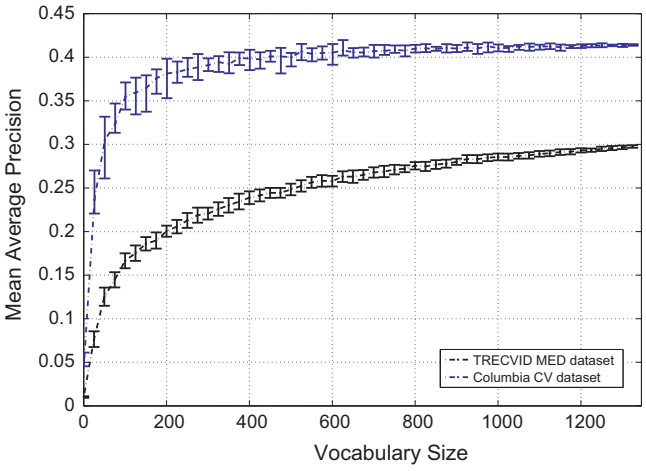



**Fig. 5.** Experiment 1: Increasing the vocabulary size improves the event recognition performance. This improvement is especially prevalent for small vocabularies containing less than 200 concept detectors.

When we analyze individual event recognition results, we observe a few cases exist where a single type vocabulary outperforms the joint type because of the tight connection between the event description and specific concepts. For example, using a single type vocabulary made of *animals* only, we achieve a higher average precision for *"feeding animal"*, *"grooming animal"* and *"dog show"* events in comparison to a joint type vocabulary on the TRECVID MED dataset. Similarly, the *"Ice Skating"* and *"Skiing"* events from the Columbia CV dataset are recognized better by the action concepts than by the joint vocabulary. Nevertheless, joint type vocabularies do better than single type vocabularies on average. Therefore, we consider joint type vocabularies more suited for general purpose event recognition. The performance difference between the single type and joint type vocabularies varies per concept type. For some types, like *animal*, the difference is substantial (0.158 vs. 0.239 and 0.310 vs. 0.265 on the TRECVID MED and the Columbia CV dataset respectively), while for others, like *action*, it is almost negligible (0.067 vs. 0.076 and 0.197 vs. 0.217 on the TRECVID MED and the Columbia CV dataset respectively). We attribute the performance difference to at least two reasons. First, our concept detectors are trained on a global image level, so they contain a considerable amount of contextual information. Consequently, some single types may contain a wide sample of contextual information including 'semantic overlap' from other concept types. The *action* pool, for example, may contain action detectors in varying scenes using various objects. Second, when creating several concept detectors for a similar type, it is likely that the detectors will be correlated, especially for the less diverse types, *e.g., People* and *Animal*. To clarify this observation we plot the correlation between concept detectors within a concept type in Fig. 6. As shown in this figure, the highly correlated concepts tend to belong to the same concept type. Therefore, including too many concepts from the same type in a vocabulary, especially from the less diverse concept types like *animal* and *people*, leads to correlated concepts and should be avoided.

We recommend to make the vocabulary diverse by including concepts from various concept types and to limit the number of concepts for the less diverse types.

### 6.3. Experiment 3: Which concepts?

Tables 4 and 5 compare three types of vocabularies: specific, general and mixture. According to the MAPs on both datasets, the general vocabulary performs better than the specific vocabulary, but the mixture vocabulary is in both cases the best overall performer.

We observe that for a few events a specific vocabulary outperforms the others, *e.g., "repairing appliance"* and *"music performance"*. For these events, there are some specific and discriminative concepts available in the vocabulary. For example, the `washing machine`, `refrigerator` and `microwave` concepts for *"repairing appliance"* and `music stool`, `instrumental musician` and `acoustic guitar` concepts for *"music performance"*. While the specific concepts may be distinctive for recognizing some events, the concepts typically occur in only few videos. Hence, they are absent in most videos and do not contribute much to generic event recognition. Therefore, if the vocabulary consists of specific concepts only, it will perform well in recognizing the events relevant to those concepts, but it will perform poor for other events. In contrast to the specific concepts, general concepts occur in a large numbers of videos. Although these concepts are not discriminative individually, taking several of them together into a vocabulary makes the event recognition better than using a specific vocabulary. Since it is able to simultaneously utilize distinctive specific concepts and general concepts, the best performance is obtained when

**Table 2**
Experiment 2: Comparison of single type and joint type vocabularies for event recognition on the TRECVID MED dataset. Each column pair compares a single and joint type vocabulary. To make the vocabularies more comparable within a concept type, we force them to be of equal size. Note that the number of concept detectors (in parenthesis) varies per concept type, so comparison across concept types should be avoided. The results demonstrate that for all the six concept types, joint type vocabularies outperform single type vocabularies on average. Best results per type denoted in bold.

| Event | Concept type | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Object(670) | | Action(34) | | Scene(128) | | People(78) | | Animal(321) | | Attribute(45) | |
| | Single | Joint | Single | Joint | Single | Joint | Single | Joint | Single | Joint | Single | Joint |
| Attempting board trick | **0.368** | 0.348 | 0.056 | **0.073** | 0.115 | **0.169** | 0.065 | **0.119** | 0.120 | **0.271** | **0.082** | 0.079 |
| Feeding animal | 0.035 | **0.044** | 0.029 | **0.074** | 0.024 | **0.042** | 0.040 | **0.041** | **0.073** | 0.045 | **0.055** | 0.037 |
| Landing fish | 0.337 | **0.423** | 0.055 | **0.076** | 0.157 | **0.246** | 0.074 | **0.182** | 0.323 | **0.360** | 0.054 | **0.111** |
| Wedding ceremony | 0.493 | **0.520** | 0.054 | **0.073** | 0.139 | **0.193** | **0.141** | 0.119 | 0.162 | **0.388** | 0.040 | **0.070** |
| Working wood working project | 0.194 | **0.203** | 0.029 | **0.040** | 0.074 | **0.101** | **0.118** | 0.072 | 0.116 | **0.167** | 0.032 | **0.048** |
| Birthday party | 0.264 | **0.277** | 0.098 | **0.099** | 0.115 | **0.174** | **0.138** | 0.131 | 0.139 | **0.239** | 0.058 | **0.095** |
| Changing vehicle tire | 0.171 | **0.174** | 0.034 | **0.054** | 0.073 | **0.105** | 0.036 | **0.076** | 0.054 | **0.153** | 0.043 | **0.052** |
| Flash mob gathering | 0.471 | **0.494** | **0.257** | 0.212 | **0.349** | 0.304 | 0.321 | **0.337** | 0.415 | **0.475** | **0.273** | 0.251 |
| Getting vehicle unstuck | 0.330 | **0.362** | 0.092 | **0.138** | 0.186 | **0.268** | 0.110 | **0.217** | 0.294 | **0.338** | 0.069 | **0.154** |
| Grooming animal | 0.126 | **0.149** | 0.033 | **0.070** | 0.129 | **0.147** | 0.075 | **0.080** | **0.146** | 0.127 | **0.075** | 0.068 |
| Making sandwich | 0.178 | **0.197** | 0.023 | **0.061** | 0.116 | **0.127** | 0.050 | **0.098** | 0.070 | **0.176** | 0.029 | **0.066** |
| Parade | 0.268 | **0.304** | **0.169** | 0.119 | 0.215 | **0.219** | 0.119 | **0.182** | 0.126 | **0.275** | 0.093 | **0.141** |
| Parkour | 0.398 | **0.432** | 0.023 | **0.063** | 0.150 | **0.234** | 0.034 | **0.147** | 0.089 | **0.356** | 0.031 | **0.074** |
| Repairing appliance | 0.244 | **0.323** | 0.063 | **0.078** | 0.192 | **0.224** | 0.086 | **0.126** | 0.104 | **0.259** | **0.100** | 0.083 |
| Working sewing project | **0.295** | 0.252 | 0.048 | **0.075** | 0.129 | **0.163** | 0.107 | **0.123** | 0.194 | **0.238** | 0.021 | **0.082** |
| Attempting bike trick | 0.480 | **0.502** | **0.264** | 0.076 | **0.250** | 0.245 | 0.037 | **0.171** | 0.129 | **0.392** | 0.031 | **0.096** |
| Cleaning appliance | **0.079** | 0.064 | 0.019 | **0.039** | 0.022 | **0.049** | 0.021 | **0.045** | 0.029 | **0.058** | 0.015 | **0.035** |
| Dog show | 0.500 | **0.534** | 0.093 | **0.102** | 0.423 | **0.455** | 0.114 | **0.236** | **0.555** | 0.512 | 0.116 | **0.122** |
| Giving directions location | 0.029 | **0.031** | 0.013 | **0.027** | 0.019 | **0.025** | 0.011 | **0.021** | 0.016 | **0.029** | 0.012 | **0.021** |
| Marriage proposal | 0.069 | **0.075** | 0.016 | **0.024** | 0.030 | **0.033** | **0.027** | 0.023 | 0.018 | **0.050** | 0.010 | **0.016** |
| Renovating home | 0.179 | **0.232** | 0.011 | **0.049** | 0.071 | **0.120** | 0.019 | **0.078** | 0.085 | **0.192** | 0.016 | **0.053** |
| Rock climbing | 0.347 | **0.375** | 0.027 | **0.092** | **0.217** | 0.176 | 0.101 | **0.173** | 0.309 | **0.322** | 0.063 | **0.104** |
| Town hall meeting | 0.424 | **0.456** | 0.059 | **0.099** | **0.270** | 0.244 | 0.116 | **0.172** | 0.266 | **0.379** | **0.158** | 0.115 |
| Winning race without vehicle | 0.139 | **0.147** | **0.082** | 0.061 | 0.075 | **0.101** | 0.069 | **0.081** | 0.088 | **0.138** | **0.073** | 0.060 |
| Working metal crafts project | 0.052 | **0.054** | 0.019 | **0.032** | 0.018 | **0.033** | 0.020 | **0.029** | 0.019 | **0.038** | 0.020 | **0.024** |
| *Mean* | 0.259 | **0.279** | 0.067 | **0.076** | 0.142 | **0.168** | 0.082 | **0.123** | 0.158 | **0.239** | 0.063 | **0.082** |

**Table 3**
Experiment 2: Comparison of single type and joint type vocabularies for event recognition on the Columbia CV dataset, following the explanation in Table 2. Best results per type denoted in bold.

| Event | Concept type | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Object(670) | | Action(34) | | Scene(128) | | People(78) | | Animal(321) | | Attribute(45) | |
| | Single | Joint | Single | Joint | Single | Joint | Single | Joint | Single | Joint | Single | Joint |
| Basketball | 0.309 | **0.352** | 0.258 | **0.270** | 0.380 | **0.426** | 0.172 | **0.354** | 0.322 | **0.371** | 0.217 | **0.305** |
| Baseball | 0.159 | **0.233** | 0.151 | **0.217** | 0.180 | **0.205** | 0.172 | **0.279** | 0.161 | **0.233** | **0.144** | **0.158** |
| Soccer | 0.176 | **0.221** | 0.173 | **0.212** | **0.233** | 0.229 | 0.228 | **0.267** | 0.211 | **0.217** | 0.197 | **0.221** |
| Ice Skating | 0.275 | **0.296** | **0.248** | 0.200 | 0.204 | **0.284** | 0.264 | **0.364** | 0.168 | **0.296** | 0.128 | **0.228** |
| Skiing | 0.455 | **0.499** | **0.375** | 0.288 | 0.294 | **0.417** | 0.394 | **0.404** | 0.320 | **0.437** | 0.189 | **0.327** |
| Swimming | 0.573 | **0.597** | 0.220 | **0.256** | 0.328 | **0.405** | 0.338 | **0.346** | **0.542** | 0.516 | 0.149 | **0.288** |
| Biking | **0.234** | 0.201 | 0.106 | **0.116** | 0.183 | **0.210** | 0.109 | **0.162** | 0.128 | **0.205** | 0.096 | **0.131** |
| Graduation | 0.220 | **0.294** | 0.095 | **0.120** | 0.105 | **0.147** | 0.119 | **0.129** | 0.170 | **0.176** | 0.106 | **0.118** |
| Birthday | 0.295 | **0.303** | 0.164 | **0.228** | 0.219 | **0.242** | **0.217** | 0.202 | 0.275 | **0.286** | 0.189 | **0.193** |
| Wedding reception | 0.172 | **0.179** | 0.126 | **0.144** | **0.168** | 0.162 | **0.149** | 0.122 | 0.166 | **0.170** | **0.144** | 0.133 |
| Wedding ceremony | **0.225** | 0.276 | 0.172 | **0.195** | 0.153 | **0.239** | **0.226** | 0.175 | 0.130 | **0.268** | 0.165 | **0.194** |
| Wedding dance | 0.414 | **0.433** | 0.195 | **0.198** | 0.375 | 0.340 | 0.231 | **0.286** | 0.399 | **0.407** | 0.218 | **0.227** |
| Music performance | 0.413 | **0.418** | 0.231 | **0.240** | 0.314 | **0.338** | 0.305 | **0.316** | 0.375 | **0.387** | 0.268 | **0.280** |
| Non-music performance | 0.302 | **0.305** | 0.199 | **0.256** | 0.244 | **0.268** | 0.215 | **0.247** | 0.270 | **0.291** | 0.214 | **0.222** |
| Parade | 0.388 | **0.421** | 0.241 | **0.320** | 0.355 | **0.357** | 0.293 | **0.327** | 0.334 | **0.396** | 0.247 | **0.273** |
| *Mean* | 0.307 | **0.335** | 0.197 | **0.217** | 0.249 | **0.285** | 0.229 | **0.265** | 0.265 | **0.310** | 0.178 | **0.220** |

the vocabulary contains a mixture of both specific and general concepts.

We recommend to insert both general and specific concepts into the event recognition vocabulary.

### 6.4. Experiment 4: How accurate?

Tables 6 and 7 demonstrate the effect of training less accurate vocabulary concept detectors on event recognition performance. Comparing *100%-3SIFT-SP* and *30%-3SIFT-SP* demonstrates the effect of using less examples to train vocabulary concept detectors. It shows that training concept detectors on 30% of the available training data does not substantially degrade the event recognition performance. More specifically on the TRECVID MED dataset, the performance is degraded only by a relative 8% in terms of MAP, and on the Columbia CV dataset the event recognition performance is not degraded at all. Comparing *30%-3SIFT-SP* and *30%-SIFT-SP* demonstrates the effect of using fewer descriptor types in training the detectors. It shows that using only SIFT descriptors, rather than concatenation of SIFT, Opponent-SIFT and C-SIFT descriptors, degrades the MAP is only by a relative 4% and 5% for the TRECVID MED and the Columbia CCV datasets, respectively. Furthermore,

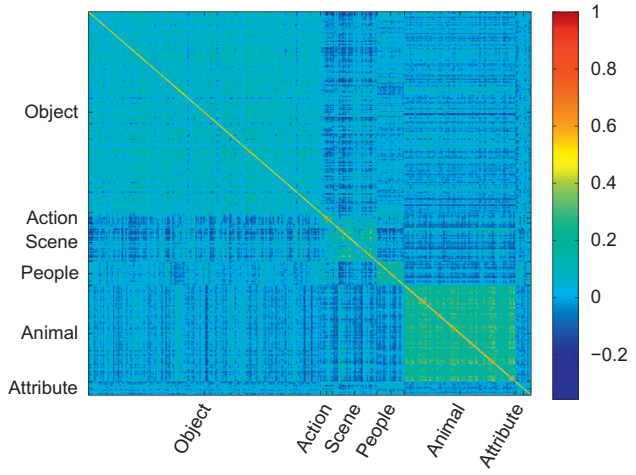



**Fig. 6.** Experiment 2: Correlation between concept detector responses appears especially within a single concept type. Including too many concepts from the same type leads to a decreased event recognition performance (matrix smoothed for better viewing, results from TRECVID MED).

**Table 4**
Experiment 3: Comparison of specific, general and mixture vocabularies for event recognition on the TRECVID MED dataset. The results demonstrate that the general vocabulary outperforms the specific vocabulary on average. The best results are obtained when the vocabulary consists of both general and specific concepts. Best results per type denoted in bold.

| Event | Specific | General | Mixture |
|---|---|---|---|
| Attempting board trick | 0.090 | 0.108 | **0.130** |
| Feeding animal | 0.041 | 0.042 | **0.045** |
| Landing fish | 0.113 | 0.107 | **0.139** |
| Wedding ceremony | 0.071 | 0.140 | **0.164** |
| Working wood working project | **0.083** | 0.065 | 0.073 |
| Birthday party | 0.078 | 0.135 | **0.138** |
| Changing vehicle tire | 0.058 | 0.062 | **0.071** |
| Flash mob gathering | 0.301 | 0.284 | **0.337** |
| Getting vehicle unstuck | 0.195 | 0.246 | **0.282** |
| Grooming animal | 0.064 | 0.079 | **0.081** |
| Making sandwich | 0.059 | 0.089 | **0.119** |
| Parade | 0.073 | **0.203** | 0.161 |
| Parkour | 0.104 | **0.226** | 0.210 |
| Repairing appliance | **0.111** | 0.098 | 0.101 |
| Working sewing project | 0.076 | 0.075 | **0.082** |
| Attempting bike trick | 0.044 | 0.080 | **0.090** |
| Cleaning appliance | **0.125** | 0.092 | 0.123 |
| Dog show | 0.219 | 0.178 | **0.230** |
| Giving directions location | 0.028 | 0.019 | **0.053** |
| Marriage proposal | 0.013 | 0.017 | **0.025** |
| Renovating home | 0.023 | 0.074 | **0.083** |
| Rock climbing | 0.178 | 0.156 | **0.194** |
| Town hall meeting | 0.064 | **0.226** | 0.158 |
| Winning race without vehicle | 0.102 | 0.102 | **0.117** |
| Working metal crafts project | **0.040** | 0.021 | 0.036 |
| *Mean* | 0.094 | 0.117 | **0.130** |

comparing *30%-SIFT-SP* and *30%-SIFT* demonstrates that including spatial pyramids in the detectors does not improve the event recognition. To summarize, the results demonstrate that the more sophisticated detectors do not substantially improve the event recognition performance.

Rather than training less sophisticated detectors we also perform the experiment with degrading the concept detector accuracies by imposing noise into their prediction scores. As expected, the results in Fig. 7 demonstrate that event recognition performance degrades by adding more noise to the concept detector predictions in the vocabulary. When the noise amount is rather small, *i.e.*, up to 30%, the event recognition remains relatively robust. For a

**Table 5**
Experiment 3: Comparison of specific, general and mixture vocabularies for event recognition on the Columbia CV dataset. Results and conclusions are similar as in Table 4. Best results per type denoted in bold.

| Event | Specific | General | Mixture |
|---|---|---|---|
| Basketball | 0.214 | 0.240 | **0.290** |
| Baseball | 0.130 | **0.273** | 0.167 |
| Soccer | 0.169 | **0.259** | 0.226 |
| Ice skating | 0.215 | 0.204 | **0.222** |
| Skiing | 0.271 | 0.195 | **0.307** |
| Swimming | 0.324 | 0.163 | **0.531** |
| Biking | 0.125 | **0.298** | 0.177 |
| Graduation | 0.097 | 0.112 | **0.191** |
| Birthday | 0.158 | **0.256** | 0.222 |
| Wedding reception | 0.112 | 0.121 | **0.129** |
| Wedding ceremony | 0.124 | 0.147 | **0.181** |
| Wedding dance | 0.263 | 0.301 | **0.316** |
| Music performance | **0.313** | 0.297 | 0.305 |
| Non-music performance | 0.224 | 0.244 | **0.255** |
| Parade | 0.376 | 0.370 | **0.384** |
| *Mean* | 0.208 | 0.232 | **0.260** |

vocabulary containing 1346 concepts, on the TRECVID MED for example, the relative performance drops by only 3% when the noise amount is 30%. When 50% noise is inserted into the concept detection results for the full vocabulary, the performance drops by 11%. It means that even if 50% of the detector predictions are distorted, the event recognition performance will be degraded by only 11%. We observe even more robust behavior against the imposed noise on the Columbia CV. Interestingly, it implies that improving the current level of concept detector accuracy has at best a limited influence on the overall event recognition performance.

What is more, improving the detector accuracies has the same effect on event recognition performance as adding more detectors to the vocabulary. If we insert 50% noise into the vocabulary made of 50 concept detectors, on the TRECVID MED for example, the event recognition performance is 0.10 in terms of MAP. We may improve the accuracy by removing the noise again, or by adding 50 more (noisy) concept detectors to the vocabulary. In both cases the event recognition performance increases to 0.13 in terms of MAP. We observe similar behavior on Columbia CV. Considering the wide availability of large amounts of training data for concept detectors [11], adding more concept detectors seems to be more straightforward than improving the detector accuracies for event recognition vocabularies.

Our experiments confirm the observation by Hauptmann et al. [9]: effective video retrieval can be achieved even when concept detector accuracies are modest, if sufficiently many concepts are combined. As a conclusion, we recommend to increase the size of the concept vocabulary rather than improving the quality of the individual detectors.

### 6.5. Experiment 5: How to normalize?

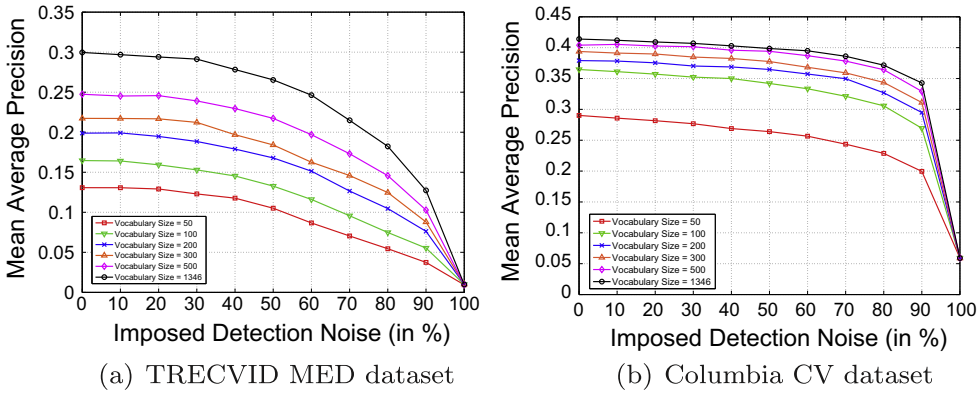
The results of this experiment, are shown in Tables 8 and 9. Both tables demonstrate that the representation obtained from un-normalized detector predictions is outperformed by all the normalized representations. More specifically on the TRECVID MED, normalizing the detector predictions by sigmoid, Z-score, and W-score normalization improves the event recognition performance, in terms of MAP, by 13%, 68% and 89%, where on Columbia CV the numbers are 27%, 164% and 169%, respectively. This substantial improvement is achieved because normalization boosts the event representation by making the predictions of different concept detectors comparable. Looking into the distribution of detector predictions, as illustrated in Fig. 8, we observe that different detectors generate different predictions distributions, which

**Table 6**
Experiment 5: Event recognition performance on the TRECVID MED dataset for four versions of vocabulary concept detectors with varying levels of robustness. The vocabularies include 346 semantic concepts trained on the TRECVID Semantic Indexing task 2012. More sophisticated concept detectors, using more training data, extra image descriptors, and spatial pyramids do not improve the event recognition performance substantially. Best results per type denoted in bold.

| Event | 100%-3SIFT-SP | 30%-3SIFT-SP | 30%-SIFT-SP | 30%-SIFT |
|---|---|---|---|---|
| Attempting board trick | 0.217 | **0.291** | 0.242 | 0.254 |
| Feeding animal | 0.045 | 0.042 | **0.075** | 0.043 |
| Landing fish | 0.231 | 0.237 | 0.309 | **0.374** |
| Wedding ceremony | **0.468** | 0.410 | 0.405 | 0.424 |
| Working wood working project | 0.118 | 0.080 | **0.131** | 0.101 |
| Birthday party | 0.119 | 0.136 | **0.148** | 0.144 |
| Changing vehicle tire | 0.156 | **0.179** | 0.088 | 0.084 |
| Flash mob gathering | 0.359 | 0.352 | 0.347 | **0.380** |
| Getting vehicle unstuck | 0.229 | 0.264 | **0.274** | 0.268 |
| Grooming animal | 0.158 | 0.116 | **0.212** | 0.133 |
| Making sandwich | **0.187** | 0.142 | 0.131 | 0.139 |
| Parade | 0.229 | **0.270** | 0.184 | 0.194 |
| Parkour | **0.437** | 0.344 | 0.317 | 0.328 |
| Repairing appliance | 0.247 | 0.206 | 0.191 | **0.266** |
| Working sewing project | 0.149 | **0.186** | 0.134 | 0.097 |
| Attempting bike trick | **0.402** | 0.379 | 0.283 | 0.298 |
| Cleaning appliance | 0.042 | **0.060** | 0.044 | 0.034 |
| Dog show | **0.417** | 0.342 | 0.401 | 0.341 |
| Giving directions location | 0.028 | **0.043** | 0.042 | 0.037 |
| Marriage proposal | 0.018 | 0.024 | **0.028** | 0.022 |
| Renovating home | 0.117 | 0.065 | 0.085 | **0.125** |
| Rock climbing | **0.271** | 0.178 | 0.175 | 0.251 |
| Town hall meeting | **0.388** | 0.265 | 0.179 | 0.217 |
| Winning race without vehicle | 0.082 | 0.064 | **0.104** | 0.043 |
| Working metal crafts project | 0.044 | **0.055** | 0.033 | 0.034 |
| *Mean* | **0.206** | 0.189 | 0.182 | 0.185 |

**Table 7**
Experiment 5: Repetition of the experiment explained in Table 6 on the Columbia CV dataset. Best results per type denoted in bold.

| Event | 100%-3SIFT-SP | 30%-3SIFT-SP | 30%-SIFT-SP | 30%-SIFT |
|---|---|---|---|---|
| Basketball | 0.469 | **0.532** | 0.490 | 0.489 |
| Baseball | 0.187 | 0.177 | 0.180 | **0.206** |
| Soccer | 0.435 | 0.467 | 0.466 | **0.499** |
| Ice skating | 0.473 | **0.500** | 0.490 | 0.456 |
| Skiing | 0.494 | **0.507** | 0.457 | 0.485 |
| Swimming | 0.527 | **0.621** | 0.593 | 0.559 |
| Biking | 0.279 | **0.296** | 0.267 | 0.247 |
| Graduation | **0.207** | 0.185 | 0.177 | 0.179 |
| Birthday | 0.304 | **0.309** | 0.280 | 0.282 |
| Wedding reception | **0.228** | 0.210 | 0.198 | 0.191 |
| Wedding ceremony | 0.201 | 0.198 | 0.190 | **0.203** |
| Wedding dance | **0.504** | 0.467 | 0.481 | 0.484 |
| Music performance | **0.310** | 0.310 | 0.303 | 0.249 |
| Non-music performance | 0.277 | **0.286** | 0.267 | 0.264 |
| Parade | **0.497** | 0.494 | 0.473 | 0.495 |
| *Mean* | 0.359 | **0.371** | 0.354 | 0.353 |




(a) TRECVID MED dataset (b) Columbia CV dataset

**Fig. 7.** Experiment 5: Event recognition performance is robust when small amounts of noise are inserted into the concept detectors of the vocabulary. The more accurate the concept detectors in a vocabulary, the higher the event recognition performance. However, adding more detectors with the same noise levels may be a more straightforward way to increase event recognition performance.

**Table 8**
Experiment 6: Comparison of different methods for normalizing concept detector predictions on the TRECVID MED dataset. Normalization improves the event recognition performance substantially. Best results per type denoted in bold.

| Event | Un-normalized | Sigmoid [39] | Z-score [12] | W-score [43] |
|---|---|---|---|---|
| Attempting board trick | 0.173 | 0.186 | 0.300 | **0.417** |
| Feeding animal | 0.052 | **0.052** | 0.031 | 0.043 |
| Landing fish | 0.151 | 0.228 | 0.336 | **0.439** |
| Wedding ceremony | 0.113 | 0.115 | 0.476 | **0.569** |
| Working wood working project | 0.161 | 0.166 | 0.150 | **0.185** |
| Birthday party | 0.174 | 0.177 | 0.321 | **0.323** |
| Changing vehicle tire | 0.075 | 0.089 | **0.224** | 0.207 |
| Flash mob gathering | 0.400 | 0.426 | 0.406 | **0.500** |
| Getting vehicle unstuck | 0.319 | 0.308 | **0.396** | 0.391 |
| Grooming animal | 0.105 | 0.124 | **0.167** | 0.166 |
| Making sandwich | 0.125 | 0.130 | 0.185 | **0.213** |
| Parade | 0.227 | 0.240 | 0.236 | **0.323** |
| Parkour | 0.100 | 0.119 | 0.399 | **0.482** |
| Repairing appliance | 0.120 | 0.127 | **0.396** | 0.376 |
| Working sewing project | 0.203 | 0.211 | 0.282 | **0.351** |
| Attempting bike trick | 0.206 | 0.280 | 0.387 | **0.494** |
| Cleaning appliance | 0.060 | 0.049 | 0.061 | **0.070** |
| Dog show | 0.353 | 0.424 | **0.569** | 0.537 |
| Giving directions location | 0.050 | 0.053 | **0.053** | 0.035 |
| Marriage proposal | 0.029 | 0.033 | 0.032 | **0.075** |
| Renovating home | 0.127 | 0.135 | **0.239** | 0.234 |
| Rock climbing | 0.280 | 0.313 | 0.330 | **0.380** |
| Town hall meeting | 0.248 | 0.370 | 0.422 | **0.483** |
| Winning race without vehicle | 0.096 | 0.133 | **0.151** | 0.112 |
| Working metal crafts project | 0.034 | 0.037 | **0.137** | 0.095 |
| *Mean* | 0.159 | 0.181 | 0.267 | **0.300** |

**Table 9**
Experiment 6: Repetition of the experiment explained in Table 8 on the Columbia CV dataset. The effect of normalization is even more prevalent. Best results per type denoted in bold.

| Event | Un-normalized | Sigmoid [39] | Z-score [12] | W-score [43] |
|---|---|---|---|---|
| Basketball | 0.208 | 0.384 | **0.573** | 0.560 |
| Baseball | 0.048 | 0.076 | 0.189 | **0.260** |
| Soccer | 0.143 | 0.241 | **0.465** | 0.439 |
| Ice Skating | 0.325 | 0.427 | 0.507 | **0.605** |
| Skiing | 0.286 | 0.446 | **0.584** | 0.574 |
| Swimming | 0.290 | 0.303 | **0.671** | 0.563 |
| Biking | 0.105 | 0.146 | 0.300 | **0.421** |
| Graduation | 0.102 | 0.082 | **0.282** | 0.273 |
| Birthday | 0.126 | 0.097 | **0.382** | 0.320 |
| Wedding reception | 0.148 | 0.117 | 0.278 | **0.307** |
| Wedding ceremony | 0.059 | 0.071 | 0.250 | **0.366** |
| Wedding dance | 0.173 | 0.128 | 0.532 | **0.563** |
| Music performance | 0.103 | 0.146 | 0.348 | **0.431** |
| Non-music performance | 0.103 | 0.157 | **0.326** | 0.222 |
| Parade | 0.142 | 0.179 | **0.522** | 0.421 |
| *Mean* | 0.157 | 0.200 | 0.414 | **0.422** |

are not directly comparable. For some detectors a prediction score might indicate absence of the concept, while for some other detectors exactly the same score might indicate concept presence. Normalizing the predictions makes the vocabulary concept detectors more comparable, which leads to a better event recognition. Comparing the performance of supervised sigmoid normalization, with unsupervised Z-score and W-score normalizations, we observe that unsupervised methods are more effective in representing events. This contradicts the common practice in the literature to rely on sigmoid normalization *e.g.,* [31,58,14]. As shown in Table 8 for the TRECVID MED dataset, using supervised score normalization we obtain an event recognition accuracy of 0.181, in terms of MAP. But with unsupervised normalization we achieve an MAP of 0.267 and 0.300 for Z-score and W-score, respectively. Similarly the supervised normalization is substantially outperformed by unsupervised normalizations on the Columbia CV dataset. The lower performance of supervised normalization is mainly caused by the fact that it assumes the distribution of concept presence on training and test data are similar. But this assumption is violated when the concept detectors are applied, as a vocabulary, on arbitrary videos that could have different concept presence distribution from the doctors training data. For example, the concept `Military Vehicle` might have a high probability of presence in its training data but it might never be present in the event videos. The difference in concept presence distributions between concept detector training data and event videos degrades the normalization performance, leading to a less effective event representation. As Table 8 shows, despite its simplicity Z-score normalization performs well in normalizing the detector outputs and achieves an event recognition accuracy of 0.267 in terms of MAP. We explain this by the observation that many concept detectors generate bell-shaped score distributions that could be modeled as Gaussian distribution. However, this Gaussian assumption is not valid for all the score distributions. Some concept detector distributions have

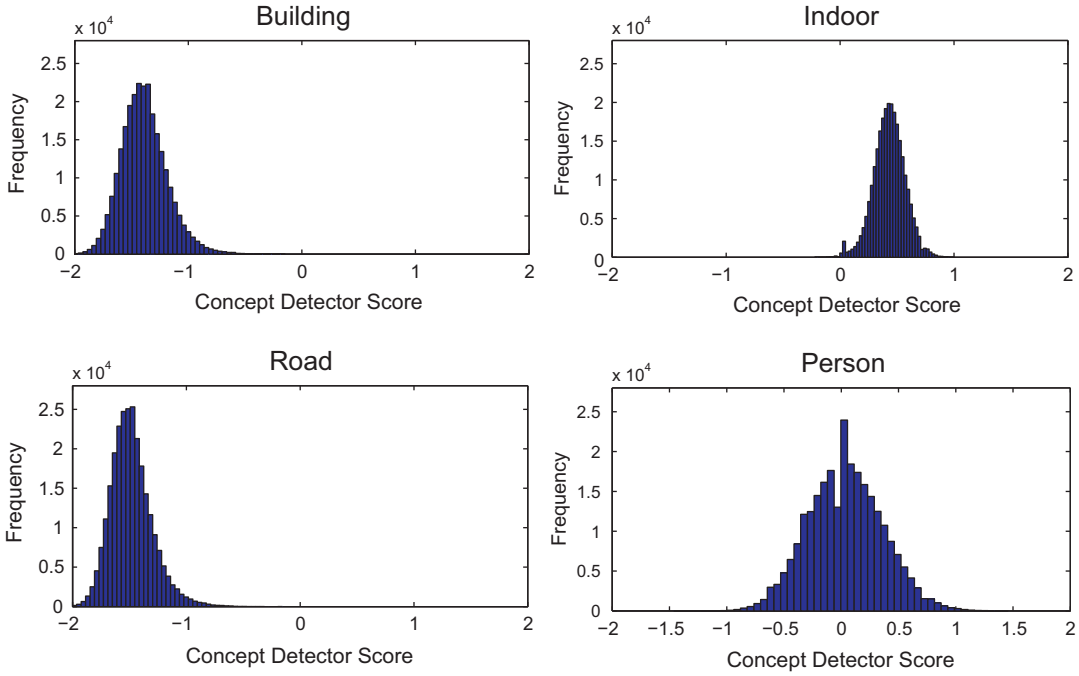



**Fig. 8.** Experiment 6: The distribution of detector predictions generated by four different concept detectors. Since the predictions have different ranges and distributions, they are not directly comparable. For example the prediction score −0.5 provides strong evidence about concept presence for `Road` and `Building`, while the same prediction indicates the concept absence for `Person` and `Indoor`. Hence, the predictions should be normalized before being used for representing events.

high skewness and some others are not even bell-shaped, which violates the Gaussian distribution assumption, so degrades the normalization effectiveness. In our experiments, the best event recognition performance is obtained after applying the unsupervised and assumption-free W-score normalization. We explain it by two reasons. First, W-score, as an unsupervised normalization, does not suffer from the possible incompatibilities between the concept distributions in the concept detector training data and the event training data. Second, W-score does not make any assumption about the overall distribution of concept detector scores, leading to better generalization.

As a conclusion, we recommend to normalize the detector predictions in a concept vocabulary, preferably by unsupervised and assumption-free normalizations.

## 7. Recommendations

In this paper we study concept vocabularies for event recognition by characterizing the concept vocabulary composition and vocabulary concept detectors. We consider five research questions related to the number, the type, the specificity, the quality and the normalization of the detectors in concept vocabularies. From the analysis of our experiments using 1346 concept detectors, two large public video datasets containing 40 events and a general event recognition pipeline, we arrive at the following five recommendations:

- **Recommendation 1:** In general, use vocabularies containing more than 200 concepts.
- **Recommendation 2:** Make the vocabulary diverse by including various concept types: *object*, *action*, *scene*, *people*, *animal* and *attributes*. However, selecting too many concepts from the same type, especially the less diverse concept types, leads to correlated concepts and should be avoided.
- **Recommendation 3:** Include both general and specific concepts into the vocabulary.
- **Recommendation 4:** Increase the size of the concept vocabulary rather than improve the quality of the individual detectors.
- **Recommendation 5:** Normalize the predictions of vocabulary concept detectors, preferably by an un-supervised and assumption-free normalization.

The recommendations may serve as guidelines to compose the appropriate concept vocabularies for future endeavors aiming for recognizing and, ultimately, explaining the semantics of complex events in video.

## Acknowledgments

This research is supported by the STW STORY project, the Dutch national program COMMIT, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government. The authors thank Dennis Koelma and Koen E.A. van de Sande for providing concept detectors.

## References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for attribute-based classification, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2013, pp. 819–826.
[2] T. Althoff, H.O. Song, T. Darrell, Detection bank: an object detection based video representation for multimedia event recognition, in: ACM International Conference on Multimedia, ACM, 2012, pp. 1065–1068.
[3] S. Ayache, G. Quénot, Video corpus annotation using active learning, in: European Conference on IR Research, Springer, 2008, pp. 187–198.
[4] N. Babaguchi, Y. Kawai, T. Kitahashi, Event based indexing of broadcasted sports video by intermodal collaboration, IEEE Trans. Multimedia 4 (1) (2002) 68–75.

[5] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, G. Serra, Event detection and recognition for semantic annotation of video, Multimedia Tools Appl. 51 (1) (2011) 279–302.
[6] A. Berg, J. Deng, S. Satheesh, H. Su, F.-F. Li, Imagenet Large Scale Visual Recognition Challenge 2011. <http://www.image-net.org/challenges/LSVRC/2011/>.
[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[8] A. Habibian, K.E.A. van de Sande, C.G.M. Snoek, Recommendations for video event recognition using concept vocabularies, in: ACM International Conference on Multimedia Retrieval, ACM, 2013, pp. 89–96.
[9] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, H. Wactlar, Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news, IEEE Trans. Multimedia 9 (5) (2007) 958–966.
[10] W. Hu, N. Xie, L. Li, X. Zeng, S. Maybank, A survey on visual content-based video indexing and retrieval, IEEE Trans. Syst., Man, Cyber., Part C: Appl. Rev. 41 (6) (2011) 797–819.
[11] B. Huet, T.-S. Chua, A. Hauptmann, Large-scale multimedia data collections, IEEE Multimedia 19 (3) (2012) 12–14.
[12] A. Jain, K. Nandakumar, A. Ross, Score normalization in multimodal biometric systems, Pattern Recognit. 38 (12) (2005) 2270–2285.
[13] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 3304–3311.
[14] L. Jiang, A. Hauptmann, G. Xiang, Leveraging high-level and low-level features for multimedia event detection, in: ACM International Conference on Multimedia, ACM, 2012, pp. 449–458.
[15] Y.-G. Jiang, Super: towards real-time event recognition in internet videos, in: ACM International Conference on Multimedia Retrieval, ACM, 2012, pp. 7–14.
[16] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, M. Shah, High-level event recognition in unconstrained videos, Int. J. Multimedia Inform. Retrieval (2013) 1–29.
[17] Y.-G. Jiang, J. Yang, C.-W. Ngo, A. Hauptmann, Representations of keypoint-based semantic concept detection: a comprehensive study, IEEE Trans. Multimedia 12 (1) (2010) 42–53.
[18] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, A.C. Loui, Consumer video understanding: a benchmark database and an evaluation of human and machine performance, in: ACM International Conference on Multimedia Retrieval, ACM, 2011, pp 29–37.
[19] Y.-G. Jiang, X. Zeng, G. Ye, D. Ellis, S.-F. Chang, S. Bhattacharya, M. Shah, Columbia-UCF trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching, in: TRECVID Workshop, TRECVID, 2010.
[20] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2-3) (2005) 107–123.
[21] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.
[22] L.-J. Li, H. Su, L. Fei-Fei, E.P. Xing, Object bank: a high-level image representation for scene classification and semantic feature sparsification, in: Advances in Neural Information Processing Systems, 2010, pp. 1378–1386.
[23] H.-T. Lin, C.-J. Lin, R.C. Weng, A note on Platt's probabilistic outputs for support vector machines, Mach. Learn. 68 (3) (2007) 267–276.
[24] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, H.S. Sawhney, Video event recognition using concept attributes, in: IEEE Workshops on Applications of Computer Vision, IEEE, 2013, pp. 339–346.
[25] X. Liu, R. Troncy, B. Huet, Finding media illustrating events, in: ACM International Conference on Multimedia Retrieval, ACM, 2011, pp. 58–65.
[26] Z. Ma, Y. Yang, Y. Cai, N. Sebe, A. Hauptmann, Knowledge adaptation for ad hoc multimedia event detection with few exemplars, in: ACM International Conference on Multimedia, ACM, 2012.
[27] Z. Ma, Y. Yang, N. Sebe, K. Zheng, A. Hauptmann, Multimedia event detection using a classifier-specific intermediate representation, IEEE Trans. Multimedia (2013) 1628–1637.
[28] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, A. Hauptmann, Complex event detection via multi-source video attributes, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2013.
[29] S. Maji, A.C. Berg, J. Malik, Efficient classification for additive kernel SVMs, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 66–77.
[30] M. Mazloom, E. Gavves, K.E.A. van de Sande, C.G.M. Snoek, Searching informative concept banks for video event detection, in: ACM International Conference on Multimedia Retrieval, ACM, 2013, pp. 255–262.
[31] M. Merler, B. Huang, L. Xie, G. Hua, A. Natsev, Semantic model vectors for complex video event recognition, IEEE Trans. Multimedia 14 (1) (2012) 88–101.
[32] G. Miller et al., Wordnet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
[33] G.K. Myers, R. Nallapati, J. van Hout, S. Pancoast, R. Nevatia, C. Sun, A. Habibian, D.C. Koelma, K.E.A. van de Sande, A.W.M. Smeulders, C.G.M. Snoek, Evaluating multimedia features and fusion for example-based event detection, Mach. Vis. Appl. 25 (1) (2014) 17–32.
[34] M. Naphade, J.R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, J. Curtis, Large-scale concept ontology for multimedia, IEEE Multimedia 13 (3) (2006) 86–91.
[35] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, Multimodal feature fusion for robust event detection in web videos, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1298–1305.
[36] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J.T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al., A large-scale benchmark dataset for event recognition in surveillance video, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 3153–3160.
[37] S. Oh, S. McCloskey, I. Kim, A. Vahdat, K.J. Cannons, H. Hajimirsadeghi, G. Mori, A.A. Perera, M. Pandey, J.J. Corso, Multimedia event detection with multimodal feature fusion and temporal concept localization, Mach. Vis. Appl. (2013) 1–21.
[38] D. Oneata, M. Douze, J. Revaud, J. Schwenninger, D. Potapov, H. Wang, Z. Harchaoui, J. Verbeek, C. Schmid, R. Aly, K. McGuinness, S. Chen, N. O'Connor, K. Chatfield, O. Parkhi, R. Arandjelovic, A. Zisserman, F. Basura, T. Tuytelaars, AXES at trecvid 2012: KIS, INS, and MED, in: TRECVID Workshop, 2012.
[39] J. Platt et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Adv. Large Margin Classifiers 10 (3) (1999) 61–74.
[40] N. Rasiwasia, N. Vasconcelos, Holistic context models for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (5) (2012) 902–917.
[41] S. Sadanand, J.J. Corso, Action bank: a high-level representation of activity in video, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1234–1241.
[42] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the Fisher vector: Theory and practice, Int. J. Comput. Vis. 105 (3) (2013) 222–245.
[43] W.J. Scheirer, N. Kumar, P.N. Belhumeur, and T.E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2933–2940, 2012.
[44] A. Scherp, R. Jain, M. Kankanhalli, V. Mezaris, Modeling, detecting, and processing events in multimedia, in: ACM International Conference on Multimedia, ACM, 2010, pp. 1739–1740.
[45] S. Shalev-Shwartz, Y. Singer, N. Srebro, A. Cotter, Pegasos: primal estimated sub-gradient solver for SVM, Math. Program. 127 (1) (2011) 3–30.
[46] J.M. Shipley, T.F. Zack (Eds.), Understanding Events, Oxford Series in Visual Cognition, Oxford University Press, 2008.
[47] A.F. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and TRECVID, in: ACM International Workshop on Multimedia Information Retrieval, 2006, pp. 321–330.
[48] C.G.M. Snoek, A.W.M. Smeulders, Visual-concept search solved?, Computer 43 (6) (2010) 76–78
[49] C.G.M. Snoek, M. Worring, Concept-based video retrieval, Found. Trends Inform. Ret. 2 (4) (2008) 215–322.
[50] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, H. Sawhney, Evaluation of low-level features and their combinations for complex event detection in open source videos, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3681–3688.
[51] L. Torresani, M. Szummer, A. Fitzgibbon, Efficient object category recognition using classemes, in: European Conference on Computer Vision, 2010, pp. 776–789.
[52] A. Ulges, C. Schulze, M. Koch, T.M. Breuel, Learning automatic concept detectors from online video, Comput. Vis. Image Understand. 114 (4) (2010) 429–438.
[53] K.E.A. van de Sande, T. Gevers, C.G.M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1582–1596.
[54] J. Varadarajan, R. Emonet, J. Odobez, Bridging the past, present and future: modeling scene activities from event relationships and global rules, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2096–2103.
[55] L. Xie, H. Sundaram, M. Campbell, Event mining in multimedia streams, Proc. IEEE 96 (4) (2008) 623–647.
[56] J. Yang, A. Hauptmann, (un)Reliability of video concept detection, in: International Conference on Image and Video Retrieval, ACM, 2008, pp. 85–94.
[57] Y. Yang, M. Shah, Complex events detection using data-driven concepts, in: European Conference on Computer Vision, Springer, 2012, pp. 722–735.
[58] E. Younessian, T. Mitamura, A. Hauptmann, Multimodal knowledge-based analysis in multimedia event detection, in: ACM International Conference on Multimedia Retrieval, ACM, 2012, pp. 51–58.
[59] J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, Int. J. Comput. Vis. 73 (2) (2007) 213–238.